Estimation of excess risk from case-control data using Aalen's linear regression model

Ørnulf Borgan
Institute of Mathematics, P.O. Box 1053 Blindern,
University of Oslo, N-0316 Oslo, Norway

Bryan Langholz
Department of Preventive Medicine,
University of Southern California, School of Medicine,
1540 Alcazar Street, CHP-220, Los Angeles, California 90033, U.S.A.

September 20, 1995

Abstract

We provide methods for statistical inference in Aalen's non-parametric linear regression model (Aalen, 1989, *Statistics in Medicine* **8** 907-925) from nested case-control data. This provides the basis for estimation of excess risk as a linear function of dose and absolute risk for a given exposure history. Tests for the hypothesis that the excess rate for an exposure is identically zero are also given, and goodness-of-fit procedures based on martingale residuals are discussed. The methods are illustrated by studying excess and absolute risks associated with radon and smoking exposure from nested case-control samples from the Colorado Plateau uranium miners cohort.

1 Introduction

While relative rate models are the standard in the analysis of epidemiologic data, and are indeed well suited for investigation of questions related to etiology of disease, excess risk models also have important uses. First, there may be exposures which truely act on an additive scale. Second, excess risk is an important measure in terms of public health, quantifying the cost, in terms of morbidity or mortality associated with exposure. Thus, simple summaries, such as excess risk per unit dose, are useful in quantifying the disease consequences of exposure.

A model that yields such quantities is Aalen's linear regression model, where the hazard function at time t for an individual with covariates $\mathbf{z}(t) = (z_1(t), \ldots, z_p(t))$ is given by

$$\alpha(t; \mathbf{z}(t)) = \alpha_0(t) + \alpha_1(t) z_1(t) + \cdots + \alpha_p(t) z_p(t). \tag{1}$$

[0] *Key words:* Excess risk; Linear hazards model; Martingale residuals; Nested case-control studies; Nonparametric methods; Survival analysis.

[0] *Abbreviated title:* Excess risk estimation in case-control studies.

The "parameters" are functions of time with $\alpha_0(t)$ the baseline hazard corresponding to $\mathbf{z}(t) = \mathbf{0}$ for all t, and $\alpha_j(t)$ the excess rate at time t per unit increase in $z_j(t)$. Let $A_j(t) = \int_0^t \alpha_j(u)\mathrm{d}u$. Throughout we will assume that the disease is rare so that these cumulative rates are good approximations to risks (Breslow and Day, 1987, Section 2.2) and will consider them to be risk measures for the remainder of this paper. Then for time-fixed covariates, $A_j(t) - A_j(s)$ is the excess risk per unit z_j between time s and t after controlling for the other covariates in the model. Further for time-varying covariates, the excess risk associated with a particular covariate history $z_j(u)$ will generally be well approximated by the simple linear expression $[A_j(t) - A_j(s)] \times \overline{z}_j$ where $\overline{z}_j$ is the average value of $z_j(u)$ over the time interval. Another quantity we will consider is $\int_s^t \alpha(u; \mathbf{z}^0(u))\mathrm{d}u$, the absolute risk of disease between s and t associated with a particular covariate history $\mathbf{z}^0(u)$.

Aalen (1980, 1989) provides estimators of the excess and absolute risks based on the $A_j(t)$ when covariate information is available for all individuals in a cohort. However, when covariate information is expensive to obtain for all cohort subjects, or as a method to reduce computational burden for large cohorts, a nested case-control sample (Thomas, 1977) is often drawn. Each case is matched to a few controls sampled from the risk set at the case's failure time resulting in a study group that is much smaller than the full cohort. Relative risks, as parameters in the Cox proportional hazards model (Cox, 1972), can be estimated from the nested case-control data using partial likelihood methods that are a natural extension of the full cohort analysis (Oakes, 1981; Borgan, Goldstein and Langholz, 1995). In this paper, we show how to estimate excess and absolute risks, as functions of the $A_j(t)$ in the Aalen linear model, from nested case-control data based on a natural extension of the methods of analysis developed for the full cohort. This requires the number of subjects at risk at each failure time in addition to the standard case-control data used in the estimation of relative risk. Tests for the hypothesis that the excess rate for an exposure is identically zero are also given, and goodness-of-fit procedures based on martingale residuals are discussed. The methods are illustrated by an excess risk analysis of lung cancer due to radon and smoking exposure using nested case-control samples from a cohort of uranium miners. In the main body of the paper, we have avoided altogether the use of counting process and martingale theory. Formal arguments and derivations using this this machinery are given in an Appendix.

Because the Aalen model is non-parametric, rather than semi-parametric as the Cox model, the number of parameter functions which can be estimated is strictly limited by the number of matched controls per case. In connection with the worked example of Section 5, however, we suggest on heuristic grounds a modification that may give useful results based on only one or two controls per case.

2 The estimator

We consider a cohort of n individuals, and let $\mathbf{z}_i(t) = (z_{i1}(t), \ldots, z_{ip}(t))$ be the covariates of the ith individual at time t. Individuals are allowed to enter and leave the population under study, and we write $\mathcal{R}(t)$ for the risk set and $n(t)$ for the number at risk at time t . Further we let $t_1 < t_2 < \cdots$ be the times when failures are observed, and denote by i_j the index of the individual failing at t_j. For nested case-control sampling, the sampled

risk set at t_j, $\widetilde{\mathcal{R}}(t_j)$, consists of the case i_j and $m-1$ controls, randomly sampled without replacement from $\mathcal{R}(t_j) \setminus \{i_j\}$. A basic assumption below is that censoring as well as sampling is independent in the sense that the additional knowledge of which individuals have been censored or sampled as controls before any time t do not alter the intensities of failure at t.

Then an estimator of $\mathbf{A}(t) = (A_0(t), A_1(t), \ldots, A_p(t))^\mathsf{T}$ may be given as follows. Define the row vectors

$$\widetilde{\mathbf{Y}}_i(t) = (1, z_{i1}(t), \ldots, z_{ip}(t)) \times \frac{n(t)}{m}, \tag{2}$$

and introduce, for each failure time t_j, the $m \times (p+1)$ matrix $\widetilde{\mathbf{Y}}(t_j)$ with rows $\widetilde{\mathbf{Y}}_i(t_j)$; $i \in \widetilde{\mathcal{R}}(t_j)$; with the first row, say, corresponding to the failure. Then the estimator for $\mathbf{A}(t)$ takes the form

$$\widetilde{\mathbf{A}}(t) = \sum_{t_j \le t} \widetilde{\mathbf{X}}(t_j)\tilde{\mathbf{e}}. \tag{3}$$

Here $\widetilde{\mathbf{X}}(t_j)$ is a generalized inverse of $\widetilde{\mathbf{Y}}(t_j)$, and $\tilde{\mathbf{e}}$ is the m dimensional column vector consisting of zeros except for a leading one corresponding to the failure. We will use the least squares generalized inverse $\widetilde{\mathbf{X}}(t_j) = [\widetilde{\mathbf{Y}}(t_j)^\mathsf{T}\widetilde{\mathbf{Y}}(t_j)]^{-1}\widetilde{\mathbf{Y}}(t_j)^\mathsf{T}$ throughout. Another possibility would have been to use a weighted least squares generalized inverse (Huffer and McKeague, 1991).

The estimator (3) is of the same form as the one for cohort data (Aalen, 1989). However for case-control data, the contribution of each subject, including the case, is weighted by the inverse of the proportion sampled from the risk set (cf. (2)). By an argument along the lines of Borgan and Langholz (1993) and Borgan, Goldstein and Langholz (1995, Section 4), it follows that $\widetilde{\mathbf{A}}(\cdot) - \mathbf{A}(\cdot)$ is almost a (vector-valued) martingale, a fact which is very useful in the study of the statistical properties of (3); details are given in the Appendix. In particular, an estimator of the covariance matrix of $\widetilde{\mathbf{A}}(t)$ is given by

$$\widetilde{\mathbf{\Sigma}}(t) = \sum_{t_j \le t} \widetilde{\mathbf{X}}(t_j)\mathrm{diag}(\tilde{\mathbf{e}})\widetilde{\mathbf{X}}(t_j)^\mathsf{T}, \tag{4}$$

where $\mathrm{diag}(\tilde{\mathbf{e}})$ is the diagonal matrix with the elements of $\tilde{\mathbf{e}}$ in the diagonal, i.e. the $(p+1) \times (p+1)$ matrix with the upper left hand entry equal one and the rest equal zero.

Large sample results of (3) may be derived by combining the arguments for Aalen's linear model for cohort data (Huffer and McKeague, 1991; Andersen, Borgan, Gill and Keiding, 1993, Section VII.4.2) with those of Borgan, Goldstein and Langholz (1995) for the Cox model based on case-control data. The result is that $\widetilde{\mathbf{A}}(\cdot) - \mathbf{A}(\cdot)$, properly normalized, asymptotically is distributed as a normal process with independent increments. We will not go into details here, however.

An important technical problem to consider is when a matrix $\widetilde{\mathbf{Y}}(t_j)$ is not of full rank. This may be a common problem for nested case-control data because the number of controls is typically small. In this case the contribution to (3) from time t_j cannot be estimated. If there are few such timepoints we let the corresponding $\widetilde{\mathbf{X}}(t_j)$ be matrices with all entries equal to zero, essentially skipping these failures. If there are many, we

suggest "pooling" controls over failure times. This technique is illustrated in the example of Section 5.

Finally, consider the absolute risk, $\int_s^t \alpha(u;\mathbf{z}^0(u))\mathrm{d}u$, associated with a given covariate history $\mathbf{z}^0(u) = (z_1^0(u),\ldots,z_p^0(u))$ over the time interval from s to t. Introducing the row vector $\mathbf{Z}^0(u) = (1, z_1^0(u),\ldots,z_p^0(u))$, this may be estimated by $\sum \mathbf{Z}^0(t_j)\Delta\widetilde{\mathbf{A}}(t_j)$ where $\Delta\widetilde{\mathbf{A}}(t_j) = \widetilde{\mathbf{X}}(t_j)\tilde{\mathbf{e}}$; cf. (3). The variance estimator takes the form $\sum \mathbf{Z}^0(t_j)\Delta\widetilde{\mathbf{\Sigma}}(t_j)\mathbf{Z}^0(t_j)^\mathsf{T}$ where $\Delta\widetilde{\mathbf{\Sigma}}(t_j) = \widetilde{\mathbf{X}}(t_j)\mathrm{diag}(\tilde{\mathbf{e}})\widetilde{\mathbf{X}}(t_j)^\mathsf{T}$; cf. (4). In both cases the sum is over all failure times t_j between s and t.

3 Hypothesis testing

In the previous section we saw how the estimator (3) for case-control data has the same form as the estimator for the full cohort. In a similar manner, the results of Aalen (1980, 1989) on hypothesis testing may be extended to case-control data.

Let us consider tests for the hypothesis that a covariate has no effect on the excess risk. To be specific, assume that we want to test the hypothesis

$$H_0 : \alpha_q(t) = 0 \text{ for all } t,$$

for some $q \geq 1$. Test statistics for H_0 may be based on

$$\widetilde{U}_q(t) = \sum_{t_j \leq t} L_q(t_j)\Delta\widetilde{A}_q(t_j), \tag{5}$$

with $\Delta\widetilde{A}_q(t_j)$ the q-th element of $\Delta\widetilde{\mathbf{A}}(t_j) = \widetilde{\mathbf{X}}(t_j)\tilde{\mathbf{e}}$. A useful choice of the (predictable) weights is to let $L_q(t_j)$ be the resiprocal of the qth diagonal element of the matrix $[\widetilde{\mathbf{Y}}(t_j)^\mathsf{T}\widetilde{\mathbf{Y}}(t_j)]^{-1}$. We will stick to this choice in Section 5. Under the null hypothesis (5) is a martingale (cf. Appendix). In particular it has expected value zero. An estimator for its variance is

$$\widetilde{V}_q(t) = \sum_{t_j \leq t} \widetilde{L}_q^2(t_j)\Delta\widetilde{\mathbf{\Sigma}}_{qq}(t_j), \tag{6}$$

where $\Delta\widetilde{\mathbf{\Sigma}}_{qq}(t_j)$ is element (q,q) of $\Delta\widetilde{\mathbf{\Sigma}}(t_j) = \widetilde{\mathbf{X}}(t_j)\mathrm{diag}(\tilde{\mathbf{e}})\widetilde{\mathbf{X}}(t_j)^\mathsf{T}$.

For situations where the alternative to H_0 is that a covariate has a positive (or negative) effect on the excess risk throughout the time span considered, a test may be based on

$$\widetilde{Z}_q(\tau) = \frac{\widetilde{U}_q(\tau)}{\sqrt{\widetilde{V}_q(\tau)}} \tag{7}$$

for a suitably choosen (large) τ. Under the null hypothesis this test statistic will be approximately standard normally distributed. On the other hand, for alternatives where the effect of a covariate is reversed as time passes, it is more useful to apply the maximal deviation statistic $\sup_{t_1 \leq t \leq t_2} |U_j(t)|/\sqrt{V_j(t)}$ over some interval $[t_1, t_2]$. Under H_0, this is asymptotically distributed as the supremum of a standardized Brownian Bridge. The

upper α-percentile d_α may therefore be found (approximately) by solving the non-linear equation

$$\frac{4\phi(d_\alpha)}{d_\alpha} + \phi(d_\alpha)\left(d_\alpha - \frac{1}{d_\alpha}\right)\log\left(\frac{\widetilde{V}_j(t_2)}{\widetilde{V}_j(t_1)}\right) = \alpha, \tag{8}$$

where $\phi(d) = (2\pi)^{-1/2}\exp(-\frac{1}{2}d^2)$ is the standard normal density (e.g. Andersen, Borgan, Gill and Keiding, 1993, Section VI.1.3 and Proposition V.4.1.c).

4 Goodness-of-fit plots

Aalen (1993) showed, for cohort data, how a usefull diagnostic tool is to plot against time the martingale residual processes aggregated over k strata. These strata may, e.g., be obtained by stratification according to the values of one or two numeric covariates. The aggregated martingale residual processes then record the observed minus (estimated) expected number of failures within each of the strata as a function of time.

Formally, for case-control data, such martingale residual process plots may be specified as follows. For each t_j, let $\mathbf{K}(t_j)$ be a $k \times m$ matix of zeros and ones, where each row indicates the individuals in $\widetilde{\mathcal{R}}(t_j)$ belonging to a particular stratum. Assume, for an illustration, that $k = 4$ and $m = 6$, i.e. that we have 4 strata and 5 controls per case. Then, if at some failure time t_j, the case belongs to stratum 3, the first control to stratum 1, the second control to stratum 4, the third and fourth controls to stratum 2 and the last control to stratum 3, one has

$$\mathbf{K}(t_j) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

The martingale residual process plots are then obtained by plotting the k components of

$$\widetilde{\mathbf{M}}^K_{\mathrm{res}}(t) = \sum_{t_j \le t} \mathbf{K}(t_j)\left[\mathbf{I}_m - \widetilde{\mathbf{Y}}(t_j)\widetilde{\mathbf{X}}(t_j)\right]\widetilde{\mathbf{e}} \tag{9}$$

against time. Here $\mathbf{I}_m$ is the the $m \times m$ identity matrix. Under the linear model (1), $\widetilde{\mathbf{M}}^K_{\mathrm{res}}(t)$ is a (vector-valued) zero mean martingale (cf. the Appendix). Thus, if the model fits well, the plots would be expected to fluctuate around the zero line.

The covariance matrix of (9) may be estimated by

$$\widetilde{\mathbf{V}}^K_{\mathrm{res}}(t) = \sum_{t_j \le t} \mathbf{K}(t_j)\left[\mathbf{I}_m - \widetilde{\mathbf{Y}}(t_j)\widetilde{\mathbf{X}}(t_j)\right]\mathrm{diag}(\widetilde{\mathbf{e}})\left[\mathbf{I}_m - \widetilde{\mathbf{Y}}(t_j)\widetilde{\mathbf{X}}(t_j)\right]^{\mathsf{T}}\mathbf{K}(t_j)^{\mathsf{T}}. \tag{10}$$

This may be used to calculate pointwise standard errors of the components of $\widetilde{\mathbf{M}}^K_{\mathrm{res}}(t)$ to see if they deviate more from the zero line than what can be explained by random variations. More formally, an ϵ-level Bonferroni type maximal deviation test may be performed as follows. The model is rejected if the maximal absolute value, over an interval $[t_1, t_2]$, of one of the components of $\widetilde{\mathbf{M}}^K_{\mathrm{res}}(t)$ divided by its standard error, exceds d_α. Here d_α is determined, separately for each component, from (8) with $\alpha = \epsilon/k$ and $\widetilde{V}_j$ replaced by the relevant component of $\widetilde{\mathbf{V}}^K_{\mathrm{res}}$.

Table 1: Test statistics of the univariate effect of radon, radon after adjustment for smoking, and interaction between radon and smoking.

Effect tested	1:2 sample	1:5 sample	1:40 sample
Radon	7.7	7.5	8.2
Radon | Smoking	–	6.9	8.3
Radon*Smoking | Radon, Smoking	–	2.6	2.6

These are standard Normal under H_0.
All P-values are < 0.01.

5 Example: The Colorado Plateau uranium miners data

The Colorado Plateau uranium miners cohort data were collected to study the effects of radon exposure and smoking on mortality rates and has been described in detail in earlier publications (e.g. Lundin, Wagoner, and Archer, 1971; Hornung and Meinhardt, 1987). We will focus on lung cancer mortality. The cohort consists of 3,347 Caucasian male miners recruited between 1950 and 1960 and was traced for mortality outcomes through December 31, 1982, by which time 258 lung cancer deaths were observed. Exposure data included radon exposure, in working level months (WLM) (Committee on the Biological Effects of Ionizing Radiation, 1988, p. 27), and smoking histories, in number of packs of cigarettes (20 cigarettes per pack) smoked per day.

For purposes of illustration, nested case-control samples were sampled from the risk sets formed with age as the time scale. The 23 tied failure times were broken randomly so that there was only one case per risk set. One sample with two, five, and 40 controls each, the latter being representative of what could be expected from the full cohort, were picked. Our main analysis questions are to quantify the excess risk per unit radon exposure and to estimate the absolute risk of lung cancer for workers with specific radon and smoking histories. For simplicity, we chose to summarize these exposure histories into cumulative radon (WLM) and cumulative packs of cigarettes lagged by two years which we denote $\mathbf{z}(t) = (R(t), S(t))$. We fitted models with $\alpha(t; \mathbf{z}(t))$ of the following form

$$\begin{aligned}
Radon: &\quad \alpha_0(t) + \alpha_R(t)R(t) && (11)\\
Radon + Smoking: &\quad \alpha_0(t) + \alpha_R(t)R(t) + \alpha_S(t)S(t) && (12)\\
Radon \times Smoking: &\quad \alpha_0(t) + \alpha_R(t)R(t) + \alpha_S(t)S(t) + \alpha_{RS}(t)R(t)S(t) && (13)
\end{aligned}$$

corresponding to radon, radon adjusted for smoking, and radon-smoking interaction models.

Table 1 gives the test statistic (7), based on all observed deaths, for testing whether parameter functions of interest in the analysis are identically equal to zero in the usual "analysis of variance" format. Thus, for instance, the test of an additive effect of radon adjusted for smoking is obtained by testing $H_0 : \alpha_R(t) = 0$ for all t in model (12). For the 1:5 and 1:40 data sets, not surprisingly, the hypotheses that radon does not increase risk is rejected. And, since there is little correlation between radon and smoking, these test

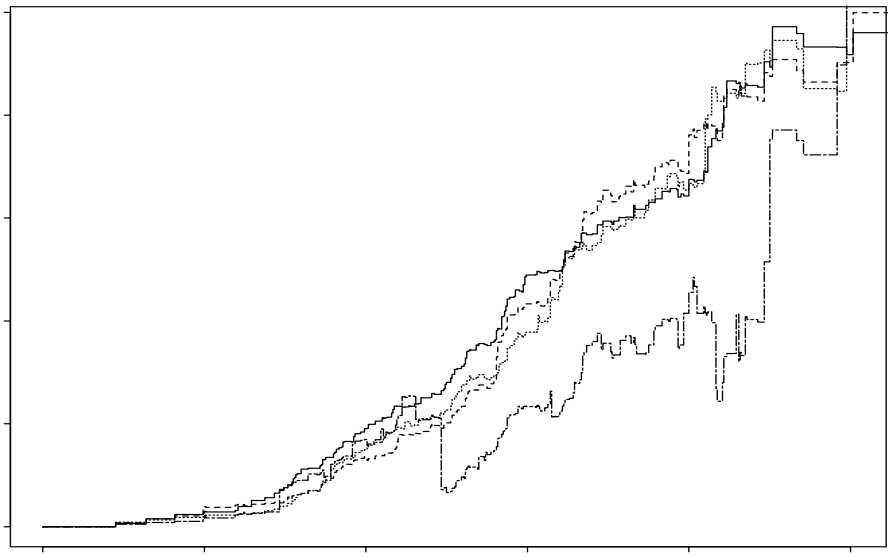

Figure 1: Estimated cumulative excess rates for radon: 1:40 sample (————-); 1:5 sample (·····); 1:2 sample (–·–·–·–·–); 1:2 pooled (– – – – – –).

statistics are not changed much after controlling for smoking. Also, each of these data sets rejects the hypothesis that there is no interaction between radon and smoking on the additive scale. With two controls, the 1:2 data set can only be used to test the radon effect in model (11) which has two parameter functions.

We estimated the cumulative excess risk curves for each of the models. While sufficient for hypothesis testing, we found that the 1:2 sample was not adequate to estimate $A_R(t) = \int_0^t \alpha_R(u)\mathrm{d}u$ in model (11). The 1:5 sample, while sensitive enough to detect the presence of the radon-smoking interaction, was not large enough to yield stable estimates in model (13).

Plots of the cumulative excess rate $A_R(t)$ for model (11) are given in Figure 1. The 1:5 sample tracks the 1:40 sample well. The 1:2 sample illustrates the nature of the instability that can occur with a small number of controls. The 1:2 curve has some large jumps due to small variation in the cumulative radon in the sampled risk sets. Such low variation sets become less likely as the number of controls increase. Intuitively, pooling controls from neighboring risk sets should improve the performance of the estimator in such situations. We did this in the 1:2 sample by pooling each case's controls with the controls of its two adjacent neighbors (in terms of age of death of the case) so that each sampled risk set had six controls. The resulting estimate, labeled "1:2 pooled," is given in Figure 1 and, indeed tracks the 1:40 sample quite closely. Variance estimation for the pooled estimator is an open problem.

Table 2 gives 10 year age-group specific excess risk for radon (unadjusted as well as adjusted for smoking) accumulated up to the lower bound of the age-group interval. Thus, for instance the (unadjusted) estimate of 6.6 for workers aged 50-59 (from the 1:5 sample)

Table 2: Excess risk (standard error) per 1000 workers per 100 WLM of cumulative radon exposure, unadjusted and adjusted for cumulative smoking by age group.

	Unadjusted			Smoking Adjusted		
Age	1:5	1:40	1:2 pooled	1:5	1:40	1:2 pooled
40-49	4.3 (0.8)	5.1 (1.0)	3.1	4.7 (1.2)	5.1 (1.0)	3.4
50-59	6.6 (1.6)	6.6 (1.3)	7.8	7.2 (1.8)	6.5 (1.3)	7.2
60-69	9.0 (3.7)	6.8 (2.2)	6.5	9.4 (4.0)	6.9 (2.3)	6.1
70-79	13.1 (10.9)	4.7 (3.1)	6.2	12.7 (8.4)	6.0 (3.0)	9.3

Table 3: Risk (standard errors), in percent, of lung cancer between ages 40 to 60 for workers with specific radon and smoking histories. Based on the fitted values for 1:5 case-control data set with the radon adjusted for smoking model.

Age	Duration	Total dose	Smoking (packs/day)		
start	(years)	(WLM)	0	1/2	1
20	20	120	0.5 (2.7)	2.2 (1.5)	4.0 (1.1)
20	20	480	4.9 (2.6)	6.7 (1.3)	8.4 (1.1)
20	20	960	10.8 (2.8)	12.6 (1.8)	14.3 (1.7)
20	40	480	3.9 (2.6)	5.6 (1.3)	7.4 (1.1)
40	20	480	2.8 (2.6)	4.5 (1.3)	6.3 (1.1)

means that the probability of lung cancer increases by 6.6 per 1000 workers for each 100 WLM of radon exposure cumulated prior to age 50. The estimates and variances are calculated simply by subtracting $\widetilde{A}_R(t)$ and $\widetilde{\Sigma}(t)$ at the upper and lower bounds of the age intervals. Estimates from the 1:5 sample vary somewhat from the 1:40, especially in the 70-79 age group, though this instability is duely reflected in the size of the standard error estimates. The 1:2 pooled estimates track the 1:40 more closely. The excess risk estimates for radon are only slightly modified when we adjust for the effect of smoking.

Next, we estimated the absolute risk of lung cancer in workers with specific radon and smoking histories based on the Radon+Smoking model (12). The smoking histories we considered were no packs, 1/2 pack, and 1 pack a day smoking starting at age 20 and continuing throughout life. For radon exposure, we considered a few starting ages, durations of exposure, and total dose combinations, all assuming a constant rate of exposure. We estimated absolute risks of lung cancer between ages 40 to 60, the results for a few combinations of parameters based on the 1:5 sample are given in Table 3. The estimated risks vary in ways one would predict. The standard errors reflect the amount of data available for estimation with the chosen parameters. There were few non-smoking cases so the standard errors among non-smokers are highest. Not withstanding the statistical significance of the interaction term in model (13), these values are all within one or two percentage points of the values predicted by the interaction model (13) using the 1:40

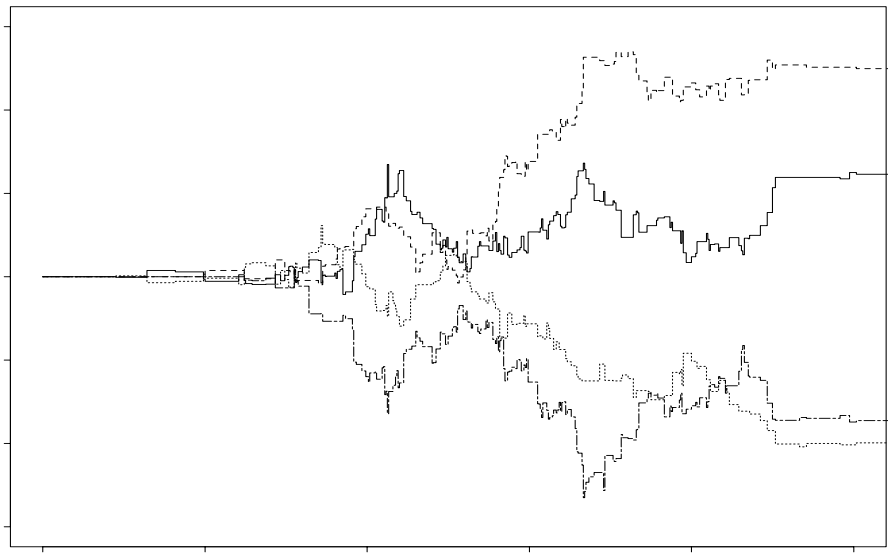

Figure 2: Martingale residual process plots for the Radon+Smoking model (12) based on the 1:5 sample. The four strata are obtained by dividing radon and smoking exposure into "low" and "high" values with cutpoints 1000 WLM for cumulative radon and 10 000 packs for cumulative smoking. The plots are: both exposures low (————); high radon and low smoking (·····); low radon and high smoking (–·–·–·–·–); both exposures high (– – – – – –).

sample.

Let us finally illustrate how the martingale residual process plots of Section 4 may be used to check the goodness-of-fit of the Radon+Smoking model (12). To this end we formed four strata by dividing both radon and smoking exposure into "low" and "high" values. As cutpoints we used 1000 WLM for cumulative radon and 10 000 packs for cumulative smoking. About three quarters of the controls in the 1:5 sample have exposure values below the cutpoint for radon while a little more than a half of the cases have exposure values which exceed it. The corresponding figures for smoking are about a half and two thirds. The resulting martingale residual process plots based on the 1:5 sample are shown in Figure 2. It is seen that the Radon+Smoking model predicts too few lung cancer deaths for the strata where both exposures are low or both are high, while it predicts too many deaths for the two strata where one of the exposures is low while the other is high. This illustrates once more that we have a positive interaction between radon and smoking exposure on the additive scale. However, for none of the four strata does the martingale residual process exceed two times its standard error. So even though Figure 2 indicates a lack of fit of the Radon+Smoking model due to the interaction between radon and smoking exposure, an omnibus type test based on the martingale residual processes will not reach statistical significance. This is in contrast to the result reported in Table 1 for the more specific test based on the Radon$\times$Smoking model (13).

6 Discussion

Methods for estimation of excess risk as a linear function of dose from case-control data have not been described previously. The methodology we present is appealing because excess risk (and standard error) associated with a given dose is computed simply by multiplying the excess risk "slope" estimate (respectively, its standard error) by the dose (as in Table 2). Absolute risk estimates associated with particular covariate histories may be easily computed and accomodate continuous, time-varying covariates. This is especially useful for occupational cohorts with persistent exposures. Alternative methods for estimation of absolute risk for time-fixed categorical covariates is described in Benichou and Gail (1995) and for continuous time-fixed covariates in Borgan and Langholz (1993).

For the ease of presentation, we have chosen to describe the methodology for simple random sampling of the controls. However, other types of risk set sampling (Langholz and Borgan, 1995; Borgan, Goldstein and Langholz, 1995) are easily accomodated by replacing $n(t)/m$ in (2) by appropriate weights as described in the Appendix.

The methods presented here are useful in nested case-control studies with multiple controls per case as is often done to reduce computation burden in large cohort studies or are occasionally conducted for the purposes of gathering covariate information on a sample of the cohort. Because of the non-parametric nature of the Aalen model, the number of parameter functions that can be estimated is strictly bounded by the number of controls sampled at each failure time. This is clearly a limitation of these methods since most nested case-control studies only collect one or two controls per case. Pooling of controls is an option in these situations, but the variance estimator (4) will underestimate the true variability. A proper variance estimator for such pooled estimators is a topic of further research.

Acknowledgements

This work was supported by National Cancer Institute grant CA14089 and Johan and Mimi Wessmann's foundation.

Appendix

We first formulate the model of Section 2 in terms of counting processes along the lines of Borgan, Goldstein and Langholz (1995). To this end introduce $\mathcal{P}$, the power set of $\{1, 2, \ldots, n\}$, i.e. the set of all subsets of $\{1, 2, \ldots, n\}$. If individual i fails at t we select the set $\mathbf{r} \in \mathcal{P}$ as our sampled risk set with probability $\pi_t(\mathbf{r}|i) = \pi_t(\mathbf{r})w_i(t)$. Here

$$w_i(t) = \frac{n(t)}{m} I(i \in \mathbf{r}), \tag{A.1}$$

and

$$\pi_t(\mathbf{r}) = \binom{n(t)}{m}^{-1} I(\mathbf{r} \subset \mathcal{R}(t), |\mathbf{r}| = m) \tag{A.2}$$

is a probability distribution over sets $\mathbf{r} \in \mathcal{P}$. Further for each $i = 1, 2, \ldots, n$ and each set $\mathbf{r} \in \mathcal{P}$, define a counting process $N_{(i,\mathbf{r})}(t)$ registering the number of times in $[0, t]$ the ith individual fails

and the sampled risk set is chosen to be $\mathbf{r}$. Then the intensity process of $N_{(i,\mathbf{r})}(t)$ is $\lambda_{(i,\mathbf{r})}(t) = I(i \in \mathcal{R}(t))\alpha(t;\mathbf{z}_i)\pi_t(\mathbf{r}|i)$. Introducing $\boldsymbol{\alpha}(t) = (\alpha_0(t), \alpha_1(t), \ldots, \alpha_p(t))^{\mathsf{T}}$ and using (1) and (2), we may write

$$\lambda_{(i,\mathbf{r})}(t) = I(i \in \mathcal{R}(t))\widetilde{\mathbf{Y}}_i(t)\boldsymbol{\alpha}(t)\pi_t(\mathbf{r}). \tag{A.3}$$

By standard counting process theory (e.g. Andersen, Borgan, Gill and Keiding, 1993, Section II.4.1) it follows that the $M_{(i,\mathbf{r})}(t) = N_{(i,\mathbf{r})}(t) - \int_0^t \lambda_{(i,\mathbf{r})}(u)\mathrm{d}u$ are orthogonal local square integrable martingales.

We then reformulate the estimator (3) in terms of counting processes. For $\mathbf{r} \in \mathcal{P}$ we introduce the $|\mathbf{r}|$ dimensional column vector $\mathbf{N}_{\mathbf{r}}(t)$ with elements $N_{(i,\mathbf{r})}(t)$, $i \in \mathbf{r}$, and define $\boldsymbol{\lambda}_{\mathbf{r}}(t)$ and $\mathbf{M}_{\mathbf{r}}(t)$ similarly. We also introduce the $|\mathbf{r}| \times (p+1)$ dimensional matrix $\widetilde{\mathbf{Y}}_{\mathbf{r}}(t)$ with rows $\widetilde{\mathbf{Y}}_i(t)I(i \in \mathcal{R}(t))$, $i \in \mathbf{r}$, cf. (2). Then the estimator (3) may be given as $\widetilde{\mathbf{A}}(t) = \sum_{\mathbf{r}\in\mathcal{P}} \widetilde{\mathbf{A}}_{\mathbf{r}}(t)$, where

$$\widetilde{\mathbf{A}}_{\mathbf{r}}(t) = \int_0^t J_{\mathbf{r}}(u)\widetilde{\mathbf{X}}_{\mathbf{r}}(u)\mathrm{d}\mathbf{N}_{\mathbf{r}}(u). \tag{A.4}$$

Here $\widetilde{\mathbf{X}}_{\mathbf{r}}(t)$ is a generalized inverse of $\widetilde{\mathbf{Y}}_{\mathbf{r}}(t)$, and $J_{\mathbf{r}}(t)$ is the indicator of $\widetilde{\mathbf{Y}}_{\mathbf{r}}(t)$ having full rank. By (A.3) we may now write $\widetilde{\mathbf{A}}_{\mathbf{r}}(t) = \mathbf{A}^*_{\mathbf{r}}(t) + \widetilde{W}_{\mathbf{r}}(t)$, where $\mathbf{A}^*_{\mathbf{r}}(t) = \int_0^t J_{\mathbf{r}}(u)\pi_u(\mathbf{r})\boldsymbol{\alpha}(u)\mathrm{d}u$ and $\widetilde{W}_{\mathbf{r}}(t) = \int_0^t J_{\mathbf{r}}(u)\widetilde{\mathbf{X}}_{\mathbf{r}}(u)\mathrm{d}\mathbf{M}_{\mathbf{r}}(u)$. Thus $\widetilde{\mathbf{A}}(t)$ equals

$$\mathbf{A}^*(t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} \{J_{\mathbf{r}}(u)\pi_u(\mathbf{r})\}\,\boldsymbol{\alpha}(u)\mathrm{d}u$$

plus the local square integrable martingale $\widetilde{W}(t) = \sum_{\mathbf{r}\in\mathcal{P}} \widetilde{W}_{\mathbf{r}}(t)$. Further if $\widetilde{\mathbf{Y}}_{\mathbf{r}}(u)$ has full rank with high probability for all $u \le t$, then $\mathbf{A}^*(t)$ is almost the same as $\mathbf{A}(t)$, and it follows that $\widetilde{\mathbf{A}}(t)$ is almost unbiased. The optional variation process of $\widetilde{W}(t)$ is $\widetilde{\boldsymbol{\Sigma}}(t) = \sum_{\mathbf{r}\in\mathcal{P}}[\widetilde{W}_{\mathbf{r}}](t)$, with

$$[\widetilde{W}_{\mathbf{r}}](t) = \int_0^t J_{\mathbf{r}}(u)\widetilde{\mathbf{X}}_{\mathbf{r}}(u)\mathrm{diag}\,(\mathrm{d}\mathbf{N}_{\mathbf{r}}(u))\,(\widetilde{\mathbf{X}}_{\mathbf{r}}(u))^{\mathsf{T}}, \tag{A.5}$$

which is seen to equal (4) and thereby providing a justification for the proposed estimator for the covariance matrix of $\widetilde{\mathbf{A}}(t)$.

The key point in the above derivations is to define the estimator $\widetilde{\mathbf{A}}(t)$, as well as the estimator of its covariance matrix, as a sum over all possible sampled risk sets. Further, the contributions from a specific set $\mathbf{r} \in \mathcal{P}$ are, except for the weights in (2), of the same form as for the full cohort; compare (A.4) and (A.5) to the formulas (7.4.5) and(7.4.9) in Andersen, Borgan, Gill and Keiding (1993). The same procedure works for hypothesis testing and martingale residual processes and gives the results summarized in Sections 3 and 4.

To be more specific, the process (5) may be written

$$\widetilde{U}_q(t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} \widetilde{L}_{\mathbf{r}q}(u)\mathrm{d}\widetilde{A}_{\mathbf{r}q}(u), \tag{A.6}$$

with $\widetilde{A}_{\mathbf{r}q}(t)$ the qth element of (A.4) and $\widetilde{L}_{\mathbf{r}q}(t)$ a predictable process. Under the hypothesis $\alpha_q(t) = 0$ for all t, (A.6) is a local square intergrable martingale. Its optional variation process equals

$$\widetilde{V}_q(t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} \widetilde{L}^2_{\mathbf{r}q}(u)\mathrm{d}[\widetilde{W}_{\mathbf{r}}](u), \tag{A.7}$$

which by (A.5) is seen to be the same as (6). This gives the basic results needed to derive the properties of the test statistics in Section 3.

For the martinagale residual processes we first define, for each $\mathbf{r} \in \mathcal{P}$,

$$\mathbf{M}_{\mathbf{r},\mathrm{res}}(t) = \mathbf{N}_{\mathbf{r}}(t) - \int_0^t \widetilde{\mathbf{Y}}_{\mathbf{r}}(u)\mathrm{d}\widetilde{\mathbf{A}}_{\mathbf{r}}(u) = \int_0^t \left(\mathbf{I}_{|\mathbf{r}|} - \widetilde{\mathbf{Y}}_{\mathbf{r}}(u)\widetilde{\mathbf{X}}_{\mathbf{r}}(u)\right) \mathrm{d}\mathbf{N}_{\mathbf{r}}(u), \tag{A.8}$$

where $\mathbf{I}_{|\mathbf{r}|}$ is the $|\mathbf{r}| \times |\mathbf{r}|$ identity matrix. Under the model (1) we may use (A.3) to see that (A.8) equals the stochastic integral

$$\mathbf{M}_{\mathbf{r},\mathrm{res}}(t) = \int_0^t \left(\mathbf{I}_{|\mathbf{r}|} - \widetilde{\mathbf{Y}}_{\mathbf{r}}(u)\widetilde{\mathbf{X}}_{\mathbf{r}}(u)\right) \mathrm{d}\mathbf{M}_{\mathbf{r}}(u).$$

This martingale has optional variation process

$$[\mathbf{M}_{\mathbf{r},\mathrm{res}}](t) = \int_0^t \left(\mathbf{I}_{|\mathbf{r}|} - \widetilde{\mathbf{Y}}_{\mathbf{r}}(u)\widetilde{\mathbf{X}}_{\mathbf{r}}(u)\right) \mathrm{diag}\left(\mathrm{d}\mathbf{N}_{\mathbf{r}}(u)\right) \left(\mathbf{I}_{|\mathbf{r}|} - \widetilde{\mathbf{Y}}_{\mathbf{r}}(u)\widetilde{\mathbf{X}}_{\mathbf{r}}(u)\right)^{\mathsf{T}}.$$

Then let $\mathbf{K}_{\mathbf{r}}(t)$ be the $k \times |\mathbf{r}|$ matrix of predictable indicator processes with (l,i)th entry equal to one if individual $i \in \mathbf{r}$ belongs to stratum l at time t. It then follows that, when model (1) is true,

$$\widetilde{\mathbf{M}}^K_{\mathrm{res}}(t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} \mathbf{K}_{\mathbf{r}}(u)\mathrm{d}\mathbf{M}_{\mathbf{r},\mathrm{res}}(u)$$

is a martingale with optional variation process

$$[\widetilde{\mathbf{M}}^K_{\mathrm{res}}](t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}} \mathbf{K}_{\mathbf{r}}(u)\mathrm{d}[\mathbf{M}_{\mathbf{r},\mathrm{res}}](u)\mathbf{K}_{\mathbf{r}}(u)^{\mathsf{T}}.$$

It is seen that these two expressions equal (9) and (10), respectively, thereby justifying the results on martingale residual processes in Section 4.

We finally note that all the results of this Appendix remain valid for the general sampling schemes of Borgan, Goldstein and Langholz (1995) provided that the weights in (2) and (A.1) and the probability distribution in (A.2) are redefined as in formula (4.2) of that paper.

References

Aalen, O. O. (1980) A model for non-parametric regression analysis of counting processes. *Springer Lect. Notes Statist.* **2**, 1-25.

Aalen, O. O. (1989) A linear regression model for the analysis of life times. *Statist. Med.* **8**, 907-925.

Aalen, O. O. (1993) Further results on the non-parametric linear regression model in survival analysis. *Statist. Med.* **12**, 1569-1588.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Verlag, New York.

Benichou, J. and Gail, M.H. (1995). Methods of inference for estimates of absolute risk derived from population-based case-control studies. *Biometrics* **51**, 182–194.

Borgan, Ø., Goldstein, L., and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Annals of Statistics*. (in press).

Borgan, Ø. and Langholz, B. (1993). Non-parametric estimation of relative mortality from nested case-control studies. *Biometrics* **49**, 593–602.

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research. Volume II – The Design and Analysis of Cohort Studies*, *IARC Scientific Publications*, Vol. 82. International Agency for Research on Cancer, Lyon.

Committee on the Biological Effects of Ionizing Radiation (1988). *Health Risks of Radon and Other Internally Deposited Alpha-Emitters, BEIR IV*, National Academy Press, Washington, D.C.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. B* **34**, 187–220.

Hornung, R. and Meinhardt, T. (1987). Quantitative risk assessment of lung cancer in U. S. uranium miners. *Health Physics*, **52**, 417–30.

Huffer, F. R. and McKeague, I. W. (1991). Weighted least squares estimation for Aalen's additive risk model. *J. Amer. Statist. Assoc.* **86**, 114-129.

Langholz, B. and Borgan, Ø. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika* **82**, 69–79.

Lundin, F., Wagoner, J., and Archer, V. (1971). Radon daughter exposure and repiratory cancer, quantitative and temporal aspects. Joint Monograph 1, U.S. Public Health Service, Washington, D.C.

Oakes, D. (1981). Survival times: Aspects of partial likelihood (with discussion). *Internat. Statist. Rev.* **49**, 235–264.

Thomas, D. C. (1977). Addendum to: Methods of cohort analysis: Appraisal by application to asbestos mining. By F. D. K. Liddell, J. C. McDonald and D. C. Thomas. *J. Roy. Statist. Soc. A* **140**, 469–491.